BIOINFORMATICS

Vol. 20 no. 15 2004, pages 2363–2369
doi:10.1093/bioinformatics/bth250

BAPS 2: enhanced possibilities for the analysis of genetic population structure

Jukka Corander*, Patrik Waldmann, Pekka Marttinen and Mikko J. Sillanpää

Rolf Nevanlinna Institute, P.O. Box 4, Fin-00014 University of Helsinki, Finland

Received on November 12, 2003; revised on January 15, 2004; accepted on January 30, 2004
Advance Access publication April 8, 2004

ABSTRACT

Summary: Bayesian statistical methods based on simulation techniques have recently been shown to provide powerful tools for the analysis of genetic population structure. We have previously developed a Markov chain Monte Carlo (MCMC) algorithm for characterizing genetically divergent groups based on molecular markers and geographical sampling design of the dataset. However, for large-scale datasets such algorithms may get stuck to local maxima in the parameter space. Therefore, we have modified our earlier algorithm to support multiple parallel MCMC chains, with enhanced features that enable considerably faster and more reliable estimation compared to the earlier version of the algorithm. We consider also a hierarchical tree representation, from which a Bayesian model-averaged structure estimate can be extracted. The algorithm is implemented in a computer program that features a user-friendly interface and built-in graphics. The enhanced features are illustrated by analyses of simulated data and an extensive human molecular dataset.

Availability: Freely available at http://www.rni.helsinki.fi/~jic/bapspage.html

Contact: jukka.corander@rni.helsinki.fi

INTRODUCTION

Recent studies of genetic population structures have applied Bayesian methods in assigning individuals or chromosomal segments into classes or clusters using multilocus molecular markers (Pritchard *et al*., 2000a; Dawson and Belkhir, 2001; Corander *et al*., 2003; Falush *et al*., 2003). Apart from the evolutionary perspective, the estimated structure can provide a useful insight into many applications, such as correcting for population stratification in association studies (Pritchard *et al*., 2000b; Satten *et al*., 2001; Sillanpää *et al*., 2001) or accounting for genetic heterogeneity (Sillanpää *et al*., 2001; Province *et al*., 2001).

Assignments in methods by Pritchard *et al*. (2000a) and Falush *et al*. (2003), are configured within a prespecified number of latent classes. Inference about the number of classes or clusters that are supported by the data is then done by repeated analyses where different fixed numbers of classes are used. Although such an approach is computationally straightforward, it has at least two drawbacks. First, as pointed out by the authors of these two papers, inference on the number of classes supported by the data is based on an *ad hoc* approximation, with a generally unknown performance. Second, in the analysis of datasets of moderate or large degree of complexity, the computational algorithm may get stuck in various parts of the parameter space in separate runs, whereby it is difficult to evaluate the correctness of the estimated distributions. Such behavior of the algorithm was reported in Rosenberg *et al*. (2002).

Dawson and Belkhir (2001) and Corander *et al*. (2003) also use molecular markers to resolve the latent genetic structure among populations, but they estimate the partition among individuals directly. This provides a natural means to infer the a priori unknown number of genetically divergent latent groups of individuals. The main differences between the approaches of Dawson and Belkhir (2001) and Corander *et al*. (2003) are in the parametric assumptions of the Bayesian model, and in that the latter lays conditions on the geographical sampling design of the individuals. It was shown in Corander *et al*. (2003) that, when applicable, the conditional approach greatly improves the statistical power to detect clusters in the data. This was illustrated with real data in Gyllenstrand and Seppä (2003). On the other hand, such an approach cannot be utilized to detect whether the data consists of a mixed or admixed sub-population within a single geographical sampling location. Therefore, we have generalized the approach of Corander *et al*. (2003) to account for such cases.

Most species are genetically structured at several levels, such as populations, demes and individuals (Weir, 1996). In our approach partitions can be formed at different levels of sampling units, where a single unit can correspond to an individual, group of individuals or even a population. Such flexibility allows the investigation of hierarchical patterns of genetic variations at levels that are of interest in any particular application.

*To whom correspondence should be addressed.

A particular advantage of using sampling units is that one can receive substantial savings in genotyping costs by applying DNA pooling techniques in connection with the method. In DNA pooling, one can determine allele frequencies for a whole pool of samples instead of individually genotyping them (e.g. Sham *et al*., 2002). Such frequencies from several different pools can then be modeled using our algorithm, as opposed to the methods that are solely restricted to individual-level analyses (Pritchard *et al*., 2000a; Dawson and Belkhir, 2001; Falush *et al*., 2003).

Our method is applicable to common types of co-dominant markers (e.g. allozymes, single-nucleotide polymorphisms, microsatellites), and to DNA haplotypes and haploid multilocus markers. Dominant markers (e.g. AFLPs) can also be used for inferring latent groups. However, estimation of evolutionary parameters from such data may not be well founded. Generally, as with other classification and clustering methods, the biological relevance of the conclusions depends on the appropriateness of the data used for the purpose of the study.

Support for running the multiple parallel Markov chain Monte Carlo (MCMC) chains introduced here enables a global-level assessment of the validity of the parameter estimates obtained from different chains. BAPS provides estimates of posterior probabilities of specific partitions of given sampling units, as well as a hierarchical-tree representation of the closeness of the sampling units, from which a Bayesian model-averaged partition can be extracted. This feature is similar to that advocated in Dawson and Belkhir (2001), and it is particularly useful when the sampling units are individuals and the molecular data are scarce, since any specific partitions have then typically very low posterior probabilities. In the software, we have also included the possibility of deriving Bayesian estimates of the fixation index (Nei, 1977) and pairwise genetic distances (Nei, 1972; Reynolds *et al*., 1983). To illustrate the new features and the potential in the BAPS software, we present analyses of simulated data and of a well-known human molecular dataset of Rosenberg *et al.* (2002).

SYSTEM AND METHODS

The target of our estimation is to provide posterior distribution of partitions among the sampling units into non-empty classes, which have non-identical allele frequency parameters over an arbitrary number of molecular marker loci. Let $S = (s_1, \dots, s_k)$ represent a partition of n sampling units into k non-empty classes. Let N_L denote the number of observed marker loci, and $N_{A(j)}$ the number of different alleles distinguishable at locus j in the data. Independently for each class s_i in S, the joint distribution of the data and parameters is proportional to the Multinomial–Dirichlet expression $\prod_{j=1}^{N_L}\prod_{l=1}^{N_{A(j)}} p_{ijl}^{n_{ijl}+\alpha_j}$, where p_{ijl} is the unknown allele frequency parameter, n_{ijl} is the observed number of copies of allele l at locus j among sampling units in s_i and α_j is the Dirichlet prior hyperparameter, chosen as $1/N_{A(j)}$. Thus, in the above we assume independence of allele frequencies between loci and Hardy–Weinberg equilibrium (HWE) within each class s_i. These assumptions are commonly used, an exception being Falush *et al.* (2003), where dependence between loci was modeled. In cases where the HWE assumption seems unjustified, our model can be used for observed genotype frequencies to account for the effect of inbreeding. The prior distribution for the parameter S is chosen to be uniform in the class of all possible partitions. The above Bayesian model used in the new BAPS 2 program for an arbitrary specified sampling unit level, is analogous to that introduced in Corander *et al.* (2003) for population-level analysis.

For small n, it is possible to use complete enumeration to obtain exactly the posterior distribution over the class $\Im$ of all possible partitions, defined as

$$p(S|\text{data}) = m(S) \Big/ \sum_{S\in\Im} m(S),$$

where $m(S)$ is the marginal likelihood (or unnormalized posterior) having the allele frequency parameters of each class s_i in S integrated out analytically (formula A1 in Corander *et al.*, 2003) according to

$$\prod_{i=1}^{k}\prod_{j=1}^{N_L}\left[\frac{\Gamma\left(\sum_l \alpha_j\right)}{\Gamma\left[\sum_l(\alpha_j + n_{ijl})\right]}\prod_{l=1}^{N_{A(j)}}\frac{\Gamma(\alpha_j + n_{ijl})}{\Gamma(\alpha_j)}\right].$$

This Bayesian goodness-of-fit measure automatically weights information across the genome, taking into account possible variation in the degree of empirical uncertainty about parameters at different loci. This is highly relevant, for instance, when the amount of missing observations vary largely over the loci. In the case of an admixed background of a particular sampling unit, e.g. an individual, the model-based averaging of genetic information will support allocation to a group having a predominant resemblance in expected allelic pattern. However, when no particular source appears in a predominant position, it is reflected in posterior uncertainty about the allocation of the particular sampling unit.

In the general case where the class of all possible partitions is too large for exhaustive enumeration, values from the posterior distribution over partitions may be generated using the Metropolis–Hastings algorithm (e.g. Robert and Casella, 1999). In the Metropolis–Hastings algorithm, a Markov chain defined in a parameter domain may be generated by random acceptance of proposal values for the next state conditional on the current state. The acceptance probability of a proposal S^* generated with probability $q(S^*|S)$, given a current value S, can be written as

$$\min\left(1, \frac{m(S^*)q(S|S^*)}{m(S)q(S^*|S)}\right),$$

where $q(S|S^{*})$ is the probability of proposing a restoration of the current state. Here, we make use of several different proposal distributions specified in the next section.

The analytical integration approach reduces considerably the computational effort needed in the MCMC, compared with the Gibbs sampling technique used by Pritchard *et al*. (2000a), where values of the allele frequency parameters are explicitly generated in each iteration. Since the model may comprise even hundreds of thousands of such parameters for large datasets, the Monte Carlo errors related to the Gibbs procedure can be of considerable magnitude.

ALGORITHM

To facilitate comparison with our earlier paper (Corander *et al*., 2003) we have listed here algorithmic improvements that are implemented in the BAPS program:

Posterior summaries

- Posterior probability estimates for different partitions are now based on analytically calculated marginal likelihoods, which reduces the amount of MCMC simulation error compared to the earlier approach that used relative frequencies of occurrences.
- We have included an ultrametric-tree representation (dendrogram) of the sampling units, from which a Bayesian model-averaged partition can be obtained.

Estimation

- The MCMC algorithm includes new move types to improve mixing within the chains, which is particularly important for clustering at the individual level.
- It is possible to apply a deterministic or stochastic partition algorithm using an arbitrary number of clusters to obtain a preferable initial configuration for the MCMC chains. This may shorten the time needed for convergence to a large extent.
- Numerical computations inside the chains have been optimized to obtain up to 50 times faster execution (single chain in the new program compared to the old algorithm).

Monitoring convergence

- Convergence of the chains can be monitored visually using built-in graphics.
- A strategy of multiple parallel MCMC chains is used, which enables a more reliable monitoring of the convergence to representative areas of the parameter space.

Data

- Possibility of using pooled DNA data.
- Support for haploid markers.

To improve the mixing properties of the simulated Markov chains, especially at individual the level, we have extended the move types that were available in the algorithm of Corander *et al*. (2003). Given the current value of any of n_c parallel Markov chains, a new value for that particular chain is proposed according to the following move types:

(1) With probability 0.5, combine two randomly chosen classes s_i, s_j.

(2) With probability 0.5, split a randomly chosen class s_i into two new classes, whose sizes are uniformly distributed between 1 and $|s_i| - 1$ (the cardinality minus one), and whose elements are randomly chosen from s_i.

(3) Move an arbitrary sampling unit from a randomly chosen class s_i with cardinality $|s_i| > 1$, into another randomly chosen class s_j.

(4) Choose one sampling unit randomly from each of two randomly chosen classes s_i and s_j, and exchange them between the classes.

In the earlier algorithm, only split and combine moves were considered, which may lead to insufficient mixing of the chains. At every iteration for each of the n_c simulated Markov chains, a random choice is made between move types 1 and 2, followed by the move types 3 and 4, upon rejection or acceptance of the first attempted move. This updating strategy is similar to that used in Dawson and Belkhir (2001) for a single chain. The proposal probabilities for the four move types simplify to the following expressions:

(1) $\binom{k}{2}^{-1}\Big/2$

(2) $\lfloor |s_i|/2 \rfloor^{-1} \binom{|s_i|}{|s_j|}^{-1}$ for $|s_j| < |s_i|/2$, and $\lfloor |s_i|/2 \rfloor^{-1} \binom{|s_i|}{|s_j|}^{-1}\Big/2$ for $|s_j| = |s_i|/2$, where s_j is one of the two new classes built from the split of s_i, with minimal cardinality $|s_j|$.

(3) $\tau(S)^{-1}(k-1)^{-1}|s_i|^{-1}$, where $\tau(S)$ is the number of classes with cardinality larger than one, and s_i is the chosen class.

(4) $\binom{k}{2}^{-1} |s_i|^{-1}|s_j|^{-1}$.

Clearly, not all move types are available at all states of the chain, which imposes trivial changes to the proposal probabilities in those cases.

Our estimate of the posterior probability $p(S|\text{data})$ of a specific partition S is given by the relative marginal likelihood $m(S)/\sum_{S\in\Im} m(S)$, where the summation is over the class of all distinct partitions found in the n_c simulated chains. Such an estimate automatically and effectively avoids giving too much weight to partitions occurring in chains that have been stuck to local maxima. This is contrary to the typically applied frequency-based approach, where the relative frequency of visits to a particular parameter configuration estimates its posterior probability. For the visual inspection of the convergence of the chains, our program shows trace plots of the marginal likelihoods for each chain.

From the above posterior distribution it is possible to calculate the marginal posterior distribution of the number of latent classes in the range $1, \ldots, n$, by summing over partitions containing an equal number of classes. Our algorithm also calculates the marginal posterior probability of the equality of allele frequencies for all pairs of sampling units. This is done by summing the posterior probabilities of the partitions having a particular pair of sampling units allocated in the same class. Thus, a measure of closeness is provided for all $\binom{n}{2}$ pairs, which can be used as a basis for building a hierarchical clustering representation of the sampling units in terms of a dendrogram (see e.g. Mardia *et al.*, 1979). Such representation was also advocated in Dawson and Belkhir (2001), and it enables a clustering of the sampling units in a Bayesian model-averaged sense, since the distance measure is obtained by averaging over the posterior distribution. This is particularly important when the data are scarce, since empirical support for any specific structure in terms of posterior probability may then be very low. However, Dawson and Belkhir (2001) used the relative frequency of co-occurrence of individuals in a single simulated chain as an estimate of the marginal posterior probability, whereas our estimate is based on the relative marginal likelihoods obtained from all chains.

EMPIRICAL RESULTS AND DISCUSSION

Simulated data analysis

The purpose of the simulated data analysis is to illustrate the usefulness of the model-averaging approach when there is a considerable degree of uncertainty in the estimated posterior distribution. First, we consider a simulated set of 100 individuals for which alleles at 50 biallelic loci ($j = 1, \ldots, 50$) were generated. The population consisted of 10 divergent groups ($i = 1, \ldots, 10$), with 10 individuals sampled from each, and with underlying allele frequency parameters (p_{ij}) determined as follows. For $i = 1, \ldots, 10$ and $j = 1, \ldots, 25$, p_{ij} were independently drawn from the uniform(0,1) distribution. Further, for $j = 26, \ldots, 50$, p_{ij} was set equal for all groups and independently drawn from the uniform(0,1) distribution for each locus. Thus, the groups have common allele frequencies at 50% of the loci, while being divergent to a random extent at the remaining half of the loci. Given these frequency parameters, we simulated a set of 100 binary vectors representing multilocus genotypes, where one of the alleles was set randomly missing at each locus for all individuals.

In the group-level analysis (exact results not shown), the different sources were rather clearly separated, and the partition with all 10 groups as isolates had the highest posterior probability (0.502). Some evidence, ranging from minor to moderate, was given for the equality of allele frequencies for population pairs 2, 3 and 5, 6. Given the small sample sizes, it was not surprising that the group structure remained somewhat uncertain. To mimic a considerable mixing of the populations, we randomly exchanged half of the individuals between the groups 1 and 2, as well as 3 and 4. The group-level analysis then erroneously allocated populations 1,2,3,4 and 6 to the same cluster [with $p(S|\text{data}) = 0.992$], while the remaining populations were kept isolated. To investigate whether BAPS could resolve the original population structure, we re-analyzed the data at individual level. In Figure 1, a dendrogram representation of the posterior closeness of the 100 individuals is given. The results are based on 100 parallel chains simulated for 100,000 iterations, which took ∼4 h on a PC with a 2.8 GHz P4 processor.

Figure 1 provides an illustration of an instance where the estimated posterior distribution has its probability mass so widely distributed that the investigation of probabilities of particular partitions is tedious. The dendrogram, however, provides a useful summary of the distribution and can be used to obtain a partition by cutting it at a suitable distance level. The distance between a pair of sampling units equals one minus the marginal posterior probability of them being allocated in the same class. When the dendrogram in Figure 1 is partitioned, say, at distance level 0.4, most of the individuals are correctly allocated with others in their groups of origins (17 out of 100 are erroneously allocated at that level).

We also checked for the consistency of our algorithm by another analysis, where the data was generated as above, except that the number of available marker loci was doubled. With the 100 simulated markers having again diverged allele frequencies at 50% of them, the original structure could be exactly resolved, both in group and individual level analyses.

Real data analysis

Recently, studies concerning the structure of human populations (Rosenberg *et al*., 2002; Bamshad *et al*., 2003) have generated a lot of discussion in several scientific forums (Cooper *et al*., 2003; Burchard *et al*., 2003; Calafell, 2003; Excoffier and Hamilton, 2003; Feldman *et al*., 2003; Haga and Venter, 2003). In the study by Rosenberg *et al*. (2002) the structure of human populations was investigated by using an extensive set of microsatellite marker genotypes at 377 loci for individuals from a worldwide sample of 52 populations.

In their admixture analysis, Rosenberg *et al*. (2002) used the Bayesian classification method of Pritchard *et al*. (2000a) for assigning genomic segments within individuals into their hidden population groups. The results displayed five very well-defined groups that seemed to correspond well to five major geographic regions excluding the additional outlier, the Kalash population. However, it was also reported that their estimation method started to converge to different solutions in separate runs when the number of population groups was specified to be higher than six. We re-analyzed their data

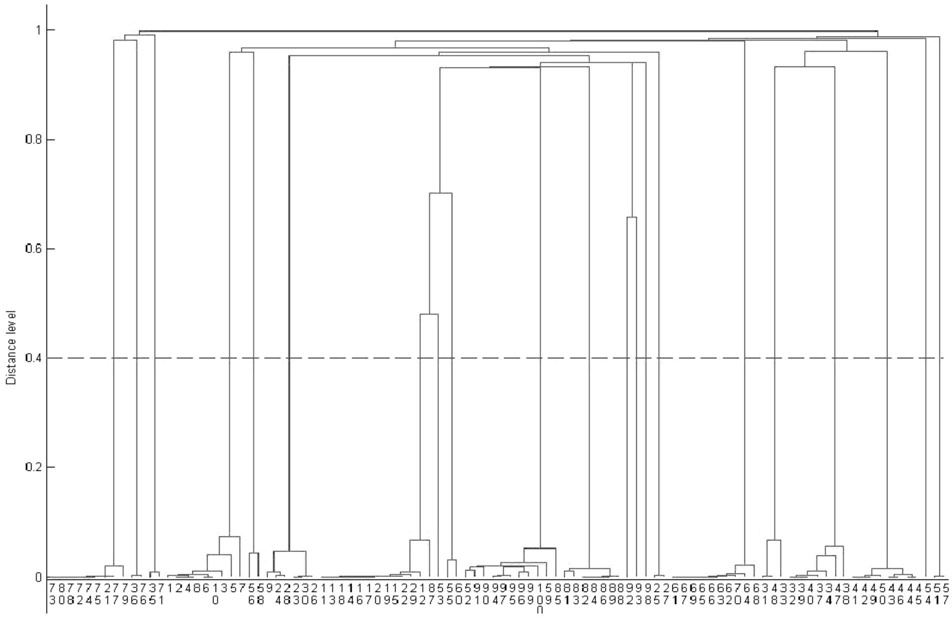

Fig. 1. A dendrogram representation of the posterior closeness of 100 simulated individuals with data from 50 biallelic marker loci, based on the complete linkage algorithm (Mardia *et al.*, 1979). The individuals are indexed in linear order according to the original 10 groups (1–10, 11–20, etc.). Individuals that are joined under the marked distance level (horizontal broken line) are allocated in the same cluster in a model-averaged partition. This figure can be viewed in colour on *Bioinformatics* online.

using BAPS, however, when comparing the results to those of Rosenberg *et al.* (2002), one should keep in mind that unlike us, Rosenberg *et al.* (2002) used an admixture rather than the mixture-based approach for their classifications.

First, to estimate the latent classes among the 52 sampled populations, we used 100 parallel chains, each of which was initialized with a partition where all original populations were treated as separate classes. In Figure 2, trace plots of the logarithms of marginal likelihoods for the chains are given for the first 5000 iterations. These clearly reveal that many chains are stuck on a considerably lower marginal likelihood level than the chain associated with the largest values. When the marginal likelihoods are converted to posterior probabilities, only a single partition (Table 1) was practically supported by the data [estimated $p(S|\text{data}) = 1.000$]. Further iterations did not change the posterior mode estimate, and in fact, Figure 2 reveals that the best chain had converged to the mode value already after 1000 iterations. The analysis with 5000 iterations took $\sim$10 h on the same computer used in the previous analyses.

Our Bayesian partition estimate coincides closely with the results of Rosenberg *et al.* (2002), however, it also shows that more groups are needed to fully represent the genetic differences at the global level. The largest divergences seem to appear between American and the remaining populations. The Africa, East Asia, Eurasia and Oceania clusters in Table 1 are identical to their counterparts obtained in Rosenberg *et al.* (2002), except for the Kalash population, which was considered as an isolate in their six-cluster solution. The populations from America were all allocated in a single cluster in the global analysis of Rosenberg *et al.* (2002), whereas our results indicate the following three divergent groups: (Colombia, Maya, Pima), (Karitiana) and (Surui).

While our results confirm most of the global level findings in Rosenberg *et al.* (2002), the differences obtained for the America region seem interesting. At least two reasonable explanations of the separation of American populations at the global level of differences can readily be given. First, the separation among the groups can be of very old origin, which seems reasonable when geographical distances and the degree of separation of populations in other continents are compared. Second, as commented by Rosenberg *et al.* (2002), genetic drift acts more rapidly in small and isolated populations, which makes the comparison of divergence times more difficult. However, since the latter argument applies to practically all populations from America included in the sample, at least the ancient separation of the two Central South American populations (Karitiana and Surui) from those of more northern origin (Colombia, Maya, Pima) seems very reasonable.

As noted earlier, events of mixture or admixture between populations may remain undetected or result in spurious structure estimates in the population-based comparison of allele frequencies. For instance, results of Rosenberg *et al.* (2002) gave a strong indication of the shared ancestry between Europe and Africa for two Middle East populations (Mozabite, Bedouin). In particular, two individuals in the Mozabite and

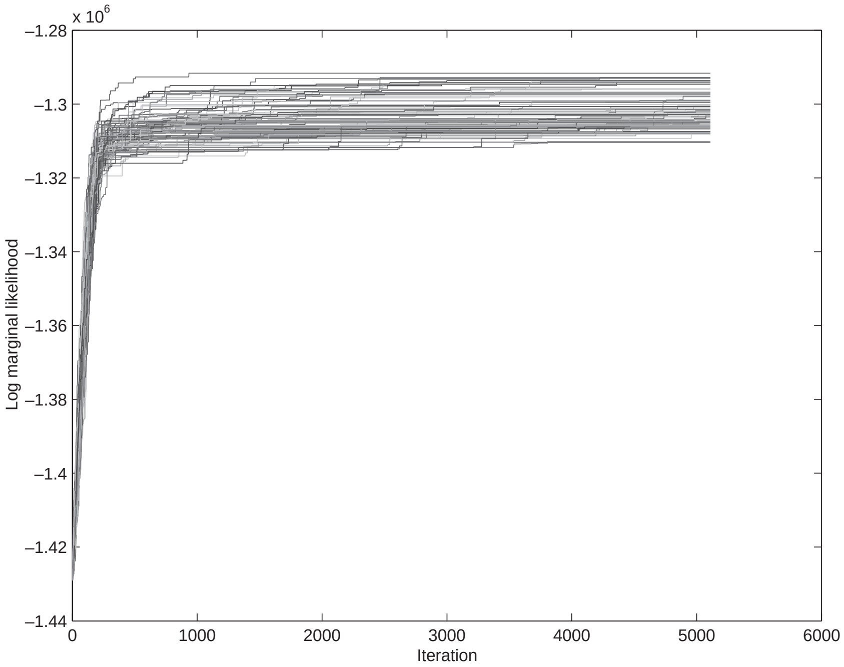

Fig. 2. Log marginal-likelihood traces of 100 MCMC chains for the first 5000 iterations in the human data analyzed at population level. This figure can be viewed in colour on *Bioinformatics* online.

Table 1. The partition of human populations with maximum posterior probability [$p(S|\text{data}) = 1.000$]

Cluster 1 (‘East Asia’):	Han, Han-N, Dai, Daur, Hezhen, Lahu, Miao, Orogen, She, Tujia, Tu, Xibo, Yi, Mongola, Naxi, Cambodian, Japanese, Yakut
Cluster 2 (‘Eurasia’):	Orcadian, Adygei, Russian, Basque, French, Italian, Sardinian, Tuscan, Mozabite, Bedouin, Druze, Palestinian, Balochi, Brahui, Burusho, Hazara, Kalash*, Makrani, Pathan, Sindhi, Uygur
Cluster 3 (‘Africa’):	Bantu, Mandenka, Yoruba, BiakaPygmy, MbutiPygmy, San
Cluster 4 (‘Oceania’):	Melanesian, Papuan
Cluster 5 (‘America’):	Colombian, Maya, Pima
Cluster 6	Kariatiana*
Cluster 7	Surui*

Populations marked with an asterisk (*) are allocated differently compared with the six cluster solution given in Rosenberg *et al*. (2002).

a single individual in the Bedouin population seemed to have strong African ancestry whereas the other members of these populations had predominant European ancestry. To investigate the possible (ad)mixture of Middle East populations we did individual-level clustering of the data from Africa, Europe and Middle East, by ignoring the known sample origins in the analysis. The data contained altogether 458 individuals from 18 different populations. The first 8 populations in the Eurasia cluster in Table 1 represent Europe, and the four populations following those represent Middle East.

In the estimation, we used 3000 iterations with 100 chains, which resulted in a single partition with estimated posterior probability equal to unity. In fact, the posterior estimate was not altered after ∼2000 iterations. The strong concentration of the posterior distribution in these two analyses of the human data (at population and individual level) illustrates how uncertainty is reduced by extensive data, even in the case of a vast parameter space. In this case, the dendrogram representation does not provide any additional information, since the posterior distribution can be compactly summarized.

The obtained individual-level partition consisted of two classes, one with the individuals from Europe and Middle East, and the other with those of African origin, apart from some exceptions. Two individuals from the Mozabite population and a single individual from the Bedouin population were allocated to the ‘Africa’ cluster, which is in agreement with the results of Rosenberg *et al*. (2002). In addition, one individual from the Biaka Pygmy population in Africa was allocated to the ‘Europe’ cluster. However, in Rosenberg *et al*. (2002) there were no significant signs of admixture for that particular population. Using the descriptive genetic distances function in BAPS, we also compared the genetic profiles of the four deviantly allocated individuals with the profiles of other individuals in their population of origin and some populations in the cluster they were assigned to. The four individuals were clearly seen more to resemble individuals in the cluster they were assigned to, than any of the other individuals in their population of origin.

Our estimated partitions reflect the divergences among the human populations that are statistically relevant at the global level from an evolutionary (drift) point of view. We would emphasize that the samples used in this study and in Rosenberg *et al*. (2002) have not been collected as a random sample within each geographical region. Instead, they have been chosen from certain distinctive human populations, based on prior information about phenotypic and cultural similarities. It is therefore likely that the correspondence between the genetic structure and the geographic regions relates more to a historical colonization pattern (mainly pre European colonial time), than to present human distribution (Feldman *et al*., 2003). Moreover, an arbitrary individual in the dataset is certainly expected to resemble genetically, individuals from the same population more than individuals from any other population, although the differences in the degree of resemblance might be very small.

In our analyses of simulated and real data we have shown that the model-based approach as implemented in BAPS, is capable of resolving genetic structure even in fairly complicated settings. However, we are currently working on some improvements on the algorithm to facilitate the mixing of the Markov chains. These, and the other new features, are worth considering in future upgrades of the BAPS software.

ACKNOWLEDGEMENTS

We would like to thank the anonymous referees for their comments and suggestions that improved the original manuscript. This work was supported by the Academy of Finland (grant no. 202324) and by the Centre of Population Genetic Analyses, University of Oulu, Finland.

REFERENCES

Bamshad,M.J., Wooding,S., Watkins,W.S., Ostler,C.T., Batzer,M.A. and Jorde,L.B. (2003) Human population genetic structure and inference of group membership. *Am. J. Hum. Genet.*, **72**, 578–589.

Burchard,E.G., Ziv,E., Coyle,N., Gomez,S.L., Tang,H., Karter,A.J., Mountain,J.L., Perez-Stable,E.J., Sheppard,D. and Risch,N. (2003) The importance of race and ethnic background in biomedical research and clinical practice. *New Engl. J. Med.*, **348**, 1170–1175.

Calafell,F. (2003) Classifying humans. *Nat. Genet.*, **33**, 435–436.

Cooper,R.S., Kaufman,J.S. and Ward,R. (2003) Race and genomics. *New Engl. J. Med.*, **348**, 1166–1170.

Corander,J., Waldmann,P. and Sillanpää,M.J. (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.

Dawson,K.J. and Belkhir,K. (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet. Res.*, **78**, 59–77.

Excoffier,L. and Hamilton,G. (2003) Comment on ‘Genetic structure of human populations’. *Science*, **300**, 1877b.

Falush,D., Stephens,M. and Pritchard,J.K. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Feldman,M.W., Lewontin,R.C. and King,M.-C. (2003) A genetic melting-pot. *Nature*, **424**, 374.

Gyllenstrand,N. and Seppä,P. (2003) Consevation genetics of the wood ant, Formica lugubris, in a fragmented landscape. *Mol. Ecol.*, **12**, 2931–2940.

Haga,S.B. and Venter,J.C. (2003) Genetics—FDA races in wrong direction. *Science*, **301**, 466.

Mardia,K.V., Kent,J.T. and Bibby,J.M. (1979) *Multivariate Analysis*. Academic Press, London.

Nei,M. (1972) Genetic distance between populations. *Am. Nat.*, **106**, 283–292.

Nei,M. (1977) F-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet.*, **41**, 225–233.

Pritchard,J.K., Stephens,M. and Donnelly,P. (2000a) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pritchard,J.K., Stephens,M., Rosenberg,N.A. and Donnelly,P. (2000b) Association mapping in structured populations. *Am. J. Hum. Genet.*, **67**, 170–181.

Province,M.A., Shannon,W.D. and Rao,D.C. (2001) Classification methods for confronting heterogeneity. *Adv. Genet.*, **42**, 273–286.

Reynolds,J., Weir,B.S. and Cockerham,C.C. (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, **105**, 767–779.

Robert,C.P. and Casella,G. (1999) *Monte Carlo Statistical Methods*. Springer, New York.

Rosenberg,N.A., Pritchard,J.K., Weber,J.L., Cann,H.M., Kidd,K.K., Zhivotovsky,L.A. and Feldmann,M.W. (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.

Satten,G.A., Flanders,W.D. and Yang,Q. (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet.*, **68**, 466–477.

Sham,P., Bader,J.S., Craig,I., O’Donovan,M. and Owen,M. (2002) DNA pooling: a tool for large-scale association studies. *Nat. Rev. Genet.*, **3**, 862–871.

Sillanpää,M.J., Kilpikari,R., Ripatti,S., Onkamo,P. and Uimari,P. (2001) Bayesian association mapping for quantitative traits in a mixture of two populations. *Genet. Epidemiol.*, **21**, S692–S699.

Weir,B.S. (1996) *Genetic Data Analysis II*. Sinauer Associates, Sunderland.